Important Announcements

- assignment 1 is due this Friday (Jan 24)
- review session by TA on Monday

- notation: random variable → upper case, X

  outcome space/set of a r.v. → caligraphic upper case, $\mathcal{X}$

  outcome of a r.v. → lower cas, x

  $$X \in \mathcal{X} = \{x_1, x_2, \dots, x_n\}$$

  $X \sim \mathbb{P}$ → "The r.v. X is sampled according to the distribution $\mathbb{P}$"

  typo: RV section $Y \in \mathcal{Y} \in \{0,1\}$

- two ways to describe a discrete random experiment
- cont. r.v. statements
- question in past lecture: infinite outcome set possible?

Motivation

Supervised Learning: Learning from a randomly sampled batch of labeled data

Dataset (tuple of tuples) $\xrightarrow[\text{(function)}]{\text{Learner}}$ **Predictor = Model = Hypothesis** (function)

$(x_{1,1}, \dots, x_{1,d})^T$

$$D = ((\vec{X}_1, Y_1), (\vec{X}_2, Y_2), \dots, (\vec{X}_n, Y_n))$$

$$D \in (\mathcal{X} \times \mathcal{Y})^n$$

$$(\vec{X}_i, Y_i) \sim \mathbb{P}_{\vec{X}, Y}$$

independent for all $i \in \{1, \dots, n\}$

D n feature-label pairs

$\mathcal{X}$ set of features

$\mathcal{Y}$ set of labels/targets

Ex: $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathbb{R}$

$$\boxed{f : \mathcal{X} \to \mathcal{Y}}$$

f a function from features to labels

Ex $f(x) = 2x + 1$, $\mathcal{X} = \mathbb{R}$

$$\boxed{\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \{f \mid f : \mathcal{X} \to \mathcal{Y}\}}$$

$\mathcal{A}$ a function from datasets to predictors

Ex: $\mathcal{A}(D) = f$ where $f : \mathcal{X} \to \mathcal{Y}$

$$f(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i \in \{1, \dots, n\} \\ 0 & \text{otherwise} \end{cases}$$

Probability

Note: Humans have a bad intuition when it comes to randomness
- Thinking Fast and Slow by: Daniel Kahneman

- Random variables
- Calculating probabilities using pmf and pdf
- Multivariate random variables
  - Conditional and marginal probabilities
- Representing random features, labels, and datasets
- Functions of random variables
- Expectation and variance

Warning: If some things seem informal, it is likely because we would need tools from Measure Theory, which we will not cover in this course.

Experiment: A process that generates an uncertain outcome

Ex: flipping a coin, rolling a dice

Outcome Space/Set: The set of all outcomes from the experiment

Ex: $\mathcal{Y} = \{0, 1\}$ Heads Tails — flipping a coin

$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ — rolling a dice

$[0, 900]$ — amount of a chemical in a wine

$\mathbb{R}$ — measurement error

The outcome space/set can be either a
1) countable set or
2) uncountable set

cardinality finite or countably infinite

cardinality is uncountably infinite

Event: A subset of the outcome space (imprecise)

Ex: Outcome space: $\mathcal{Y} = \{0, 1\}$

an outcome is a single element of the outcome set

Events: $\{0\}, \{1\}, \{0,1\} = \mathcal{Y}, \emptyset$

Ex: Outcome space: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$

Events: $\emptyset, \mathcal{X}, \{1\}, \{1,3,5\}, \{2,4,6\}, \{1,2,3\}, \ldots$

Ex: Outcome space: $[0, 900]$

Events: $\emptyset, [0,900], [0,4], [1,2] \cup [7,30], \ldots$

Probability Distribution: A function $\mathbb{P}$ defining the likelihood of each event (and satisfying certain properties)

a set containing all the events

$$\mathbb{P}: \underbrace{\text{event space/set}} \to [0,1]$$

A complicated set (σ-algebra) that we will not define

Properties: (imprecise)

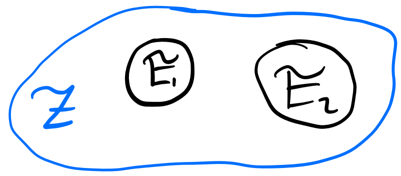
Outcome space: $\mathcal{Z}$

1. $\mathbb{P}(\mathcal{Z}) = 1$

events   disjoint

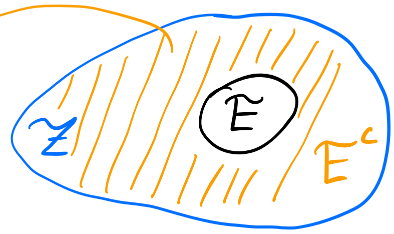
2. If $E_1 \subset \mathcal{Z}, E_2 \subset \mathcal{Z}$ and $E_1 \cap E_2 = \emptyset$, then

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$$



Ex: (of property 2.)

Events: E, E^c



$E \cap E^c = \emptyset$, $E \cup E^c = \mathcal{Z}$

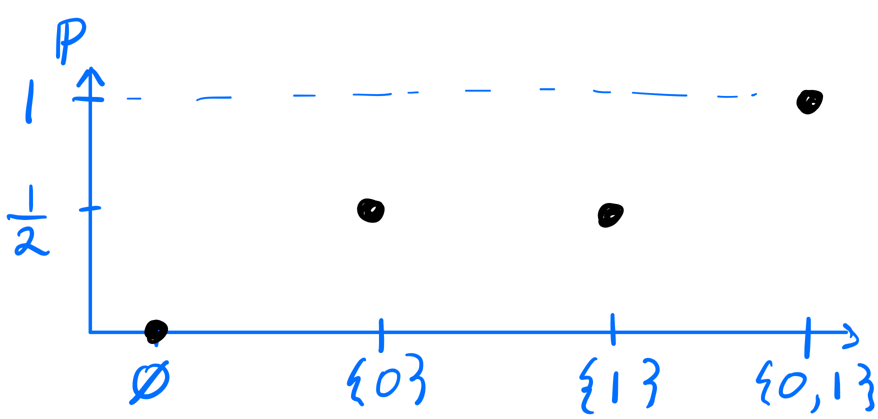
$$\underbrace{\mathbb{P}(E \cup E^c)}_{= \mathbb{P}(\mathcal{Z}) = 1} = \mathbb{P}(E) + \mathbb{P}(E^c)$$

rearranging: $\boxed{\mathbb{P}(E) = 1 - \mathbb{P}(E^c)}$

Ex: Outcome space: $\mathcal{Y} = \{0, 1\}$

$$\mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\mathcal{Y}) = 1, \ \mathbb{P}(\{0\}) = \tfrac{1}{2}, \ \mathbb{P}(\{1\}) = \tfrac{1}{2}$$

each event of an experiment is associated with a probability



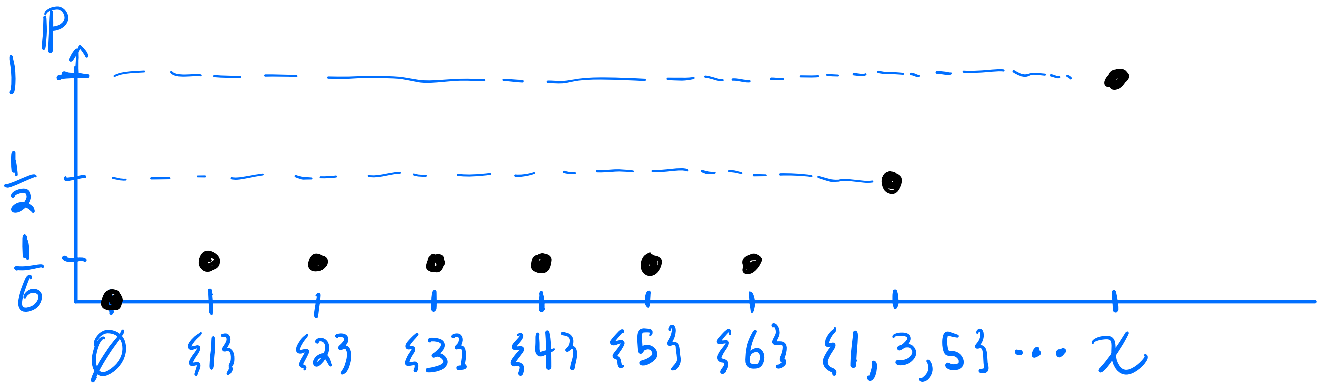
all events of experiment

Ex: Outcome space: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \dots = \mathbb{P}(\{6\}) = \frac{1}{6}$$

$$\mathbb{P}(\{1,3,5\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{5\}) = \frac{1}{2}$$

$$\mathbb{P}(\mathcal{X}) = 1$$



Random Variables

Random Variable (r.v.): (imprecise) A variable that takes a value based on the outcome of an experiment, and is associated with a probability distribution. The value can be any random outcome.

Ex: $X \in \mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ with $\mathbb{P}$ from prev. example

$Y \in \mathcal{Y} = \{0, 1\}$ with $\mathbb{P}$ from prev. example

$Z \in \{H, T\}$

A random variable is actually a function (satisfying certain properties) from one outcome space to another outcome space. Ex $X(T) = 0, X(H) = 1$.

It will not be necessary to know this for this course

Probability Distributions with r.v.

Ex: Outcome space: $\mathcal{X}$ r.v.: $X \in \mathcal{X}$

$$\mathbb{P}(\{1,3,5\}) \overset{def}{=} \mathbb{P}(\underbrace{X \in \{1,3,5\}})$$

"the r.v. X is an element of the event {1, 3, 5}"

$$\mathbb{P}(\{4,5,6\}) = \mathbb{P}(X \in \{4,5,6\}) = \mathbb{P}(X \geq 4)$$

$$\mathbb{P}(\{4\}) = \mathbb{P}(X \in \{4\}) = \mathbb{P}(X = 4)$$

Notation: $Z \sim P$ "Z is sampled according to distribution P"

Discrete r.v.: A r.v. that takes values from:

- A countable outcome space, or
- an uncountable outcome space, but there is a countable event that has probability 1

Ex: $Y \in \mathcal{Y} = \{0,1\}$, $X \in \mathcal{X} = \{1,2,3,4,5,6\}$, $Z \in \mathbb{N}$

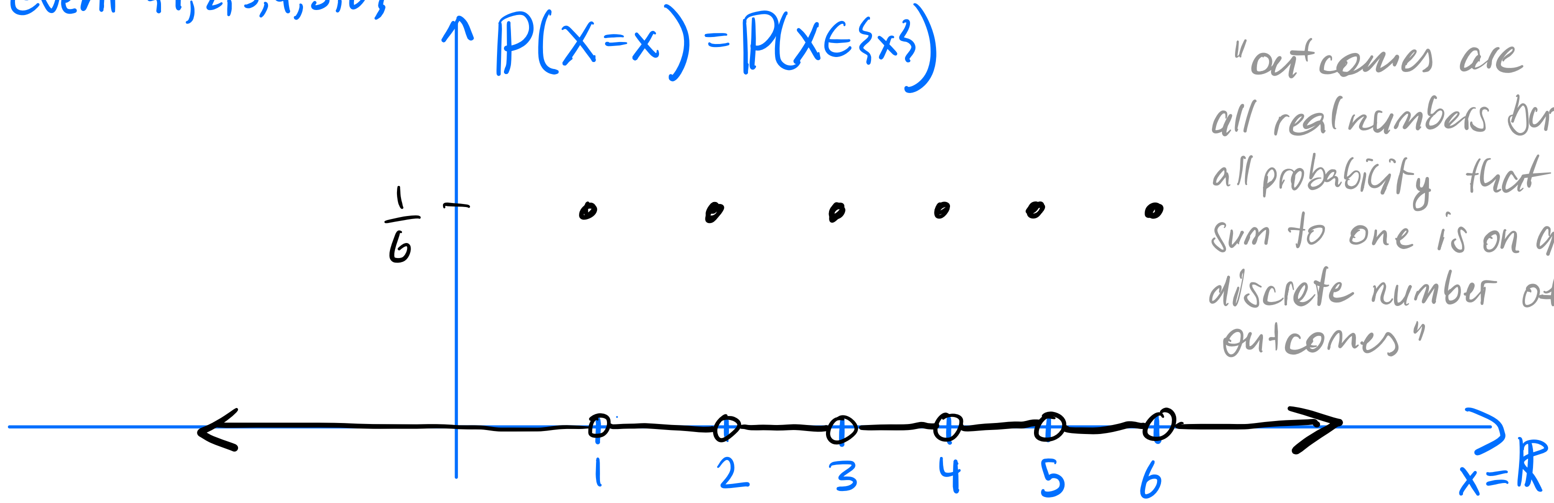
two ways to describe same discrete experiment

Ex: $X \in \mathbb{R}$ where $\mathbb{P}(X=1) = \dots = \mathbb{P}(X=6) = \frac{1}{6}$

Probability 1 for countable event $\{1,2,3,4,5,6\}$

so $\mathbb{P}(X \in \{1,2,3,4,5,6\}) = 1$

and $\mathbb{P}(\mathbb{R} \setminus \{1,2,3,4,5,6\}) = 0$



"outcomes are all real numbers but all probability that sum to one is on a discrete number of outcomes"

$\boxed{\text{Note: You can always take a r.v. defined on a countable outcome space and define it on a larger uncountable outcome space by setting the probability of the event containing all the new outcomes to zero}}$

Continuous r.v.: A r.v. that takes values from:

- an uncountable outcome space and the probability of any single outcome is zero

Ex: $Z \in [0,900]$ and $\mathbb{P}(Z=z)=\mathbb{P}(Z\in\{z\})=0$ for all $z\in[0,900]$

but $\mathbb{P}(Z\in[0,900])=1$

Ex: $Z \in \mathbb{R}$ and $\mathbb{P}(Z=z)=\mathbb{P}(Z\in\{z\})=0$ for all $z\in\mathbb{R}$

but $\mathbb{P}(Z\in\mathbb{R})=1$

Calculating Probabilities

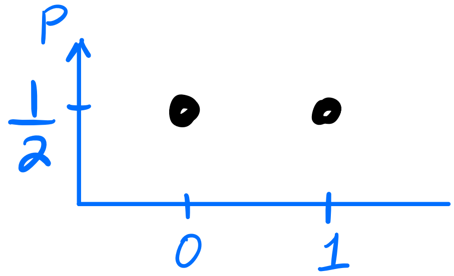
Motivation: It is hard to define the values of a probability distribution $\mathbb{P}$ for all the events

Probability Mass Function (pmf): A function $p:\mathcal{Z}\to[0,1]$ where $\mathcal{Z}$ is a discrete outcome space and $\sum_{z\in\mathcal{Z}} p(z)=1$.

The probability of an event $\mathcal{E}\subset\mathcal{Z}$ is:

$$\boxed{\mathbb{P}(Z\in\mathcal{E}) \overset{\text{def}}{=} \sum_{z\in\mathcal{E}} p(z)} \qquad \text{where } Z\in\mathcal{Z}$$

Ex: Outcome space: $\mathcal{Y}$



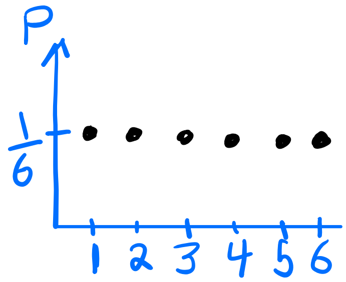
$p(0) = \frac{1}{2}$, $p(1) = \frac{1}{2}$

$$\mathbb{P}(Y \in \{0,1\}) = \sum_{y \in \{0,1\}} p(y) = p(0) + p(1) = 1$$

$$\mathbb{P}(Y=0) = \mathbb{P}(Y=1) = \mathbb{P}(Y \in \{1\}) = \sum_{y \in \{1\}} p(y) = p(1) = \frac{1}{2}$$

$$\mathbb{P}(Y \in \phi) = \sum_{y \in \phi} p(y) = 0$$

Ex: Outcome space: $\mathcal{X}$



$p(1) = p(2) = \dots = p(6) = \frac{1}{6}$

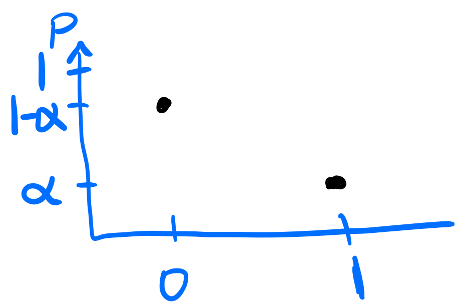
$$\mathbb{P}(X \in \{1,3,5\}) = \sum_{x \in \{1,3,5\}} p(x) = p(1) + p(3) + p(5) = \frac{1}{2}$$

Discrete Probability Distributions with special names:

Bernoulli distribution (parameter: $\alpha \in [0,1]$):

Outcome space: $\{0,1\}$

pmf: $p(1) = \alpha$, $p(0) = 1 - \alpha$



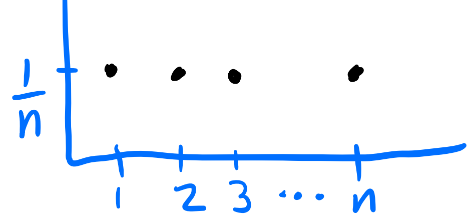
$$\boxed{\text{Distribution } \mathbb{P} = \text{Bernoulli}(\alpha)}$$

$\mathbb{P}(Z=1) = \mathbb{P}(Z \in \{1\}) = p(1) = \alpha$ $\quad Z \in \{0,1\}$ is a "Bernoulli r.v."

Discrete Uniform Distribution (parameter: n):

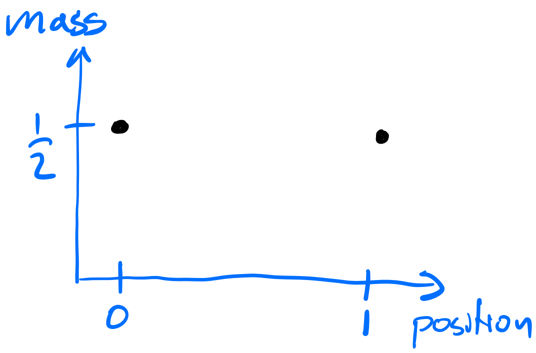
Outcome space: $\{1, 2, \dots, n\}$

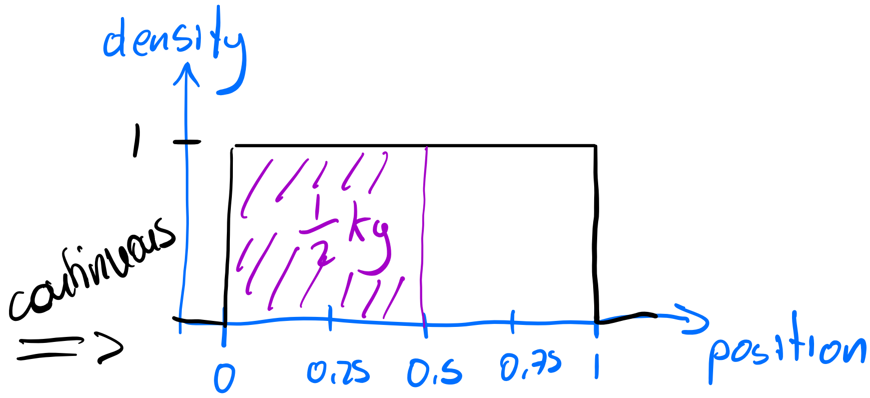
pmf: $p(1) = p(2) = \dots = p(n) = \frac{1}{n}$

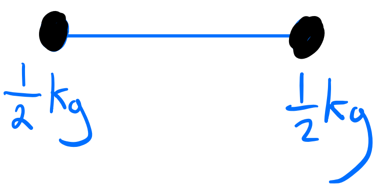


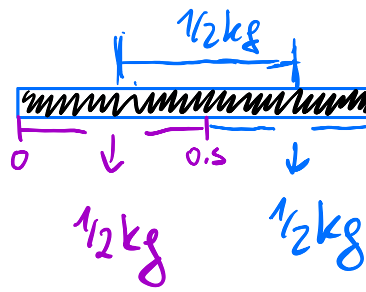
$$\boxed{\text{Distribution } \mathbb{P} = \text{Uniform}(n)}$$

Intuition with a rod in physics









Probability Density Function (pdf): a function $p: \mathcal{Z} \to [0, \infty)$ where $\mathcal{Z}$ is an uncountable outcome space and $\int_{\mathcal{Z}} p(z)\,dz = 1$

The probability of an event $E \subset \mathcal{Z}$ is:

$$\boxed{\mathbb{P}(Z \in E) \stackrel{\text{def}}{=} \int_E p(z)\,dz}$$

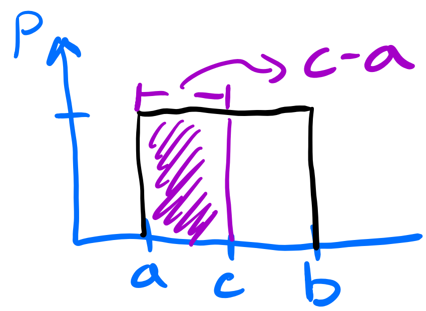
$$\boxed{\mathbb{P}(Z = z) = \mathbb{P}(Z \in \{z\}) = 0}$$ where $z \in \mathcal{Z}$

Continuous Probability Distributions with special names:

Continuous Uniform Distribution (parameters: $a \in \mathbb{R}, b \in \mathbb{R}$):

Outcome space: $[a, b]$

pdf: $p(z) =$



$$\boxed{\text{Distribution } \mathbb{P} = \text{Uniform}(a, b)}$$

$$\mathbb{P}(a \leq Z \leq c) = \mathbb{P}(Z \in [a, c]) = \int_a^c p(z)\, dz =$$

where $a \leq c \leq b$

Gaussian/Normal Distribution (parameters: $\mu \in \mathbb{R}, \sigma^2 > 0$):

standard normal dist.
$\rightarrow \mu = 0, \ \sigma^2 = 1$

Outcome space: $\mathbb{R}$

pdf: $p(z) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{1}{2\sigma^2}(z-\mu)^2\right)$

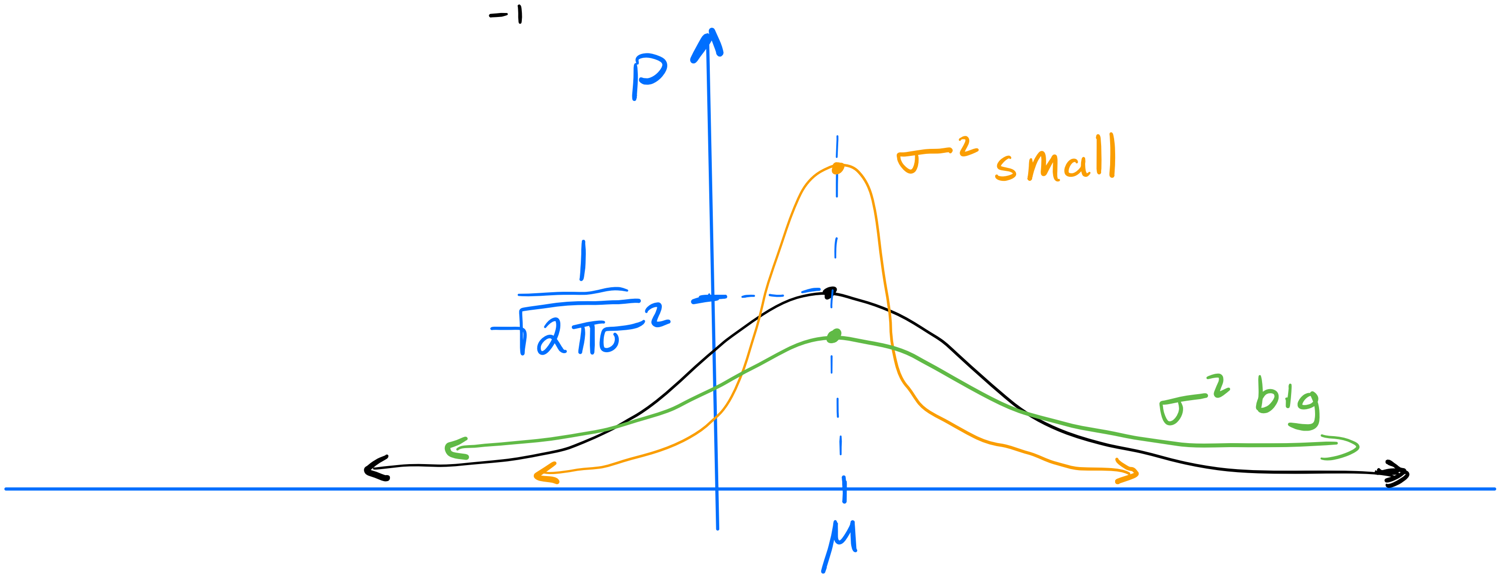
mean

variance

$$\boxed{\text{Distribution } \mathbb{P} = \mathcal{N}(\mu, \sigma^2) = \text{Gaussian}(\mu, \sigma^2)}$$

$\sigma \rightarrow$ standard deviation

Ex $\mathbb{P}(-1 \leq Z \leq 1) = \int_{-1}^{1} p(z)\, dz$

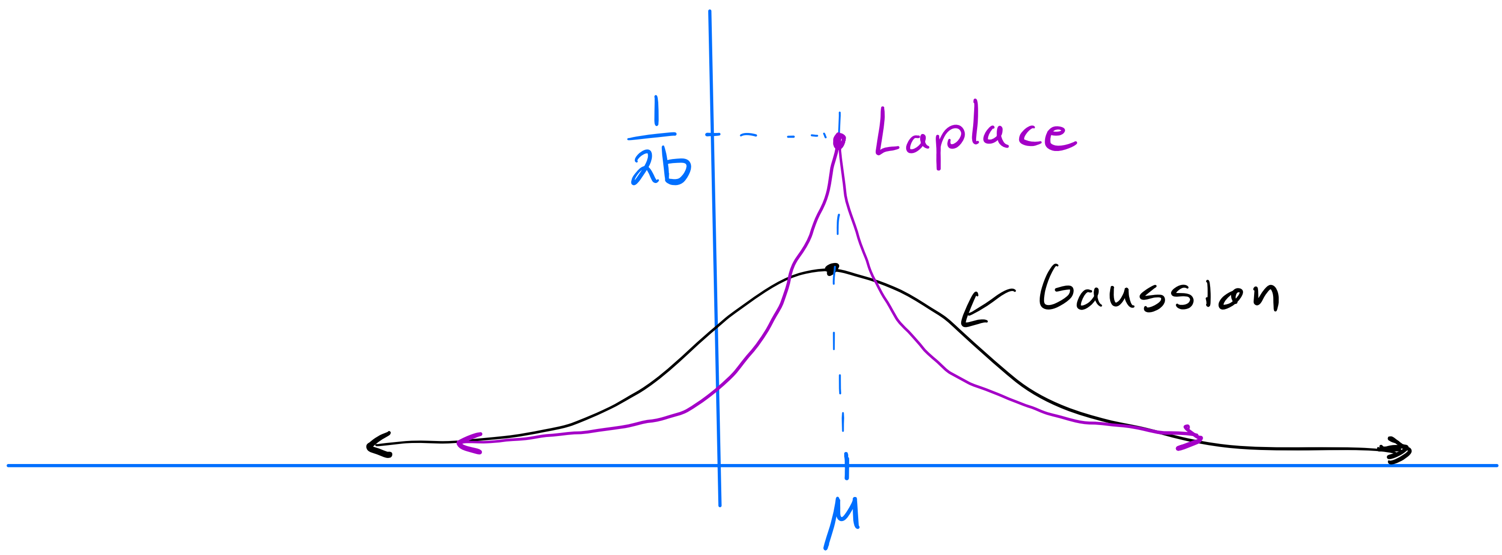


Laplace Distribution (parameters: $\mu \in \mathbb{R}, b > 0$):

Outcome space: $\mathbb{R}$

pdf: $p(z) = \frac{1}{2b} \exp\left(-\frac{1}{b}|x-\mu|\right)$

$$\boxed{\text{Distribution } \mathbb{P} = \text{Laplace}(\mu, b)}$$



Multivariate Random Variables

Motivation: To be able to talk about the probability of different types of events (outcomes of different experiments) at the same time!

Ex: The probability of getting heads and rolling a 3

The probability of a wine containing 2.5mg of one chemical and 4mg of another chemical

The probability of a house having 4 rooms and 2 washrooms and being less than 10min from a university

The probability of being young and having arthritis

Multivariate Random Variable: A tuple of more than one random variable

Ex: (Flipping 2 coins)

Outcome space: $\mathcal{X} = \{0,\overset{\text{Heads}}{1}\} \times \{0,1\} = \{(0,0),(0,1),(1,0),(1,1)\}$

r.v.: $X = (X_1, X_2) \in \mathcal{X}$

(Collecting the info of one house (ex: # of rooms, age))

Outcome space: $\mathcal{X} = \underset{\text{\# of rooms}}{\mathbb{N}} \times \overbrace{[0,\infty)}^{\text{age}}$

r.v.: $X = (X_1, X_2) \in X$

(Collecting the info of one house and its price)

Outcome space: $\mathcal{Z} = \left(\underset{\text{\# of rooms}}{\mathbb{N}} \times \overbrace{[0,\infty)}^{\text{age}}\right) \times \underbrace{[0,\infty)}_{\text{Price}}$

r.v.: $Z = (X, Y)$

$= ((X_1, X_2), Y)$

Calculating Joint Probabilities

Ex: (If you have arthritis and if you are young or old)

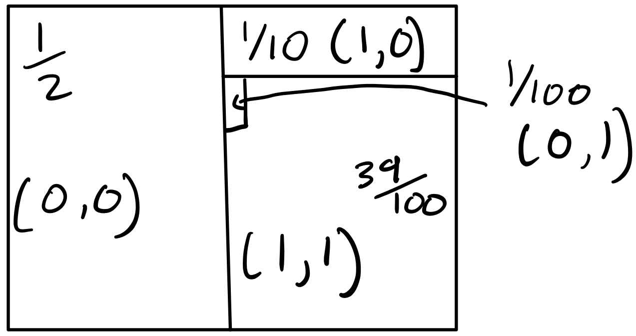
Outcome space: $\mathcal{Z} = \{\underset{\text{Young}}{0}, \overset{\text{Old}}{1}\} \times \{\underset{\text{No arthritis}}{0}, \overset{\text{Arthritis}}{1}\}$

r.v.: $Z = (X, Y) \in \mathcal{Z}$

pmf: $p: \mathcal{Z} \to [0,1]$

Not based on real data
$$\begin{cases} p((0,0)) = p(0,0) = \frac{1}{2}, & p(0,1) = \frac{1}{100} \\ p(1,0) = \frac{1}{10}, & p(1,1) = \frac{39}{100} \end{cases}$$

		Y: 0	Y: 1
X	0	1/2	1/100
	1	1/10	39/100



$$\sum_{z \in \mathcal{Z}} p(z) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) = \frac{1}{2} + \frac{1}{100} + \frac{1}{10} + \frac{39}{100} = 1$$ (for a valid pmf)

What is the probability of being young (i.e. $X=0$)?

$$E = \{(0,0),(0,1)\} = \{0\} \times \{0,1\} \subset \mathcal{Z}$$

$$\mathbb{P}(X=0, Y\in\{0,1\}) = \mathbb{P}(Z \in E) = \sum_{z \in E} p(z)$$
$$= \sum_{x\in\{0\}} \sum_{y \in\{0,1\}} p(x,y)$$
$$= p(0,0) + p(0,1)$$
$$= \frac{1}{2} + \frac{1}{100} = \frac{51}{100}$$

<u>Marginal Distribution:</u> The distribution over a subset of random variables

<u>Ex</u>: Continuing with the arthritis example

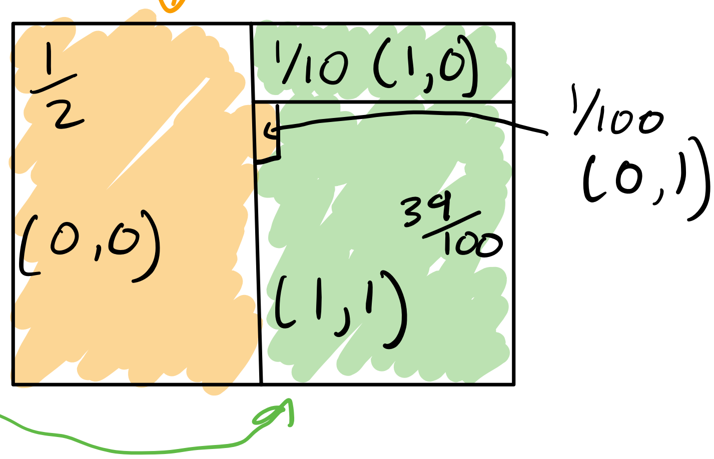
Marginal Distribution: $\mathbb{P}_X(X \in E_X)$ where $E_X \subset \mathcal{X}$

Marginal pmf: $p_X : \mathcal{X} \to [0,1]$, $p_X(x) = \sum_{y \in \mathcal{Y}} p(x,y)$

$$\mathbb{P}_X(X=0) \overset{\mathbb{P}(X \in \{0\})}{=} p_X(0) = \sum_{x \in \{0\}} \sum_{y \in \mathcal{Y}} p(0,y) = \frac{51}{100}$$

$\mathbb{P}_X(X=1)$

$$= \frac{1}{10} + \frac{39}{100} = \frac{49}{100}$$



Discrete r.v. $X_1, \dots, X_d$

$$X = (X_1, \dots, X_d) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d = \mathcal{X}, \quad p : \mathcal{X} \to [0,1]$$

$$p_{X_i} : \mathcal{X}_i \to [0,1], \quad i \in \{1, \dots, d\}$$

"Sum over every element except x_i"

Marginal pmf:

$$\boxed{p_{X_i}(x_i) \overset{\text{def}}{=} \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)}$$

$$\boxed{\mathbb{P}_{X_i}(X_i \in E_i) = \sum_{x_i \in E_i} p_{X_i}(x_i)} \quad \text{where } E_i \subset \mathcal{X}_i$$

Continuous r.v. $X_1, \dots, X_d$

$$X = (X_1, \dots, X_d) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d = \mathcal{X}, \quad p : \mathcal{X} \to [0, \infty)$$

$p_{X_i} : \mathcal{X}_i \to [0,\infty)\ ,\quad i \in \{1,\dots,d\}$

Marginal pdf: $\boxed{p_{X_i}(x_i) = \int_{\mathcal{X}_1}\cdots\int_{\mathcal{X}_{i-1}}\int_{\mathcal{X}_{i+1}}\cdots\int_{\mathcal{X}_d} p(x_1,\dots,x_d)\,dx_1\cdots dx_{i-1}\,dx_{i+1}\cdots dx_d}$

Distribution: $\boxed{\mathbb{P}_{X_i}(X_i \in E_i) = \int_{E_i} p_{X_i}(x_i)\,dx_i}$ where $E_i \subset \mathcal{X}_i$

Conditional Distribution: Probability of a r.v. given info about another r.v.

Ex: Probability that I have arthritis given I am young

Let r.v. $= Y \in \mathcal{Y},\ X \in \mathcal{X}$

Discrete Y for any $x \in \mathcal{X}$ that $p_X(x) \neq 0$

$p_{Y|X=x} : \mathcal{Y} \to [0,1],\qquad p_{Y|X=x}(y) = p_{Y|X}(y|x)$

conditional pmf: $\boxed{p_{Y|X}(y|x) \overset{\text{def}}{=} \frac{p(y,x)}{p_X(x)}}$ implies $\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x) = 1$

Distribution: $\boxed{\mathbb{P}_{Y|X}(Y \in E_Y | X = x) \overset{\text{def}}{=} \sum_{y \in E_Y} p_{Y|X}(y|x)}$ where $E_Y \subset \mathcal{Y}$

Continuous Y for any $x \in \mathcal{X}$ that $p_X(x) \neq 0$

$p_{Y|X=x} : \mathcal{Y} \to [0,\infty),\qquad p_{Y|X=x}(y) = p_{Y|X}(y|x)$

conditional pdf: $\boxed{p_{Y|X}(y|x) \stackrel{\text{def}}{=} \frac{p(y,x)}{p_X(x)}}$ implies $\int_{\mathcal{Y}} p_{Y|X}(y|x)\,dy = 1$

Distribution: $\boxed{\mathbb{P}_{Y|X}(Y \in E \mid X = x) \stackrel{\text{def}}{=} \int_{E} p_{Y|X}(y|x)\,dy}$ where $E_Y \subset \mathcal{Y}$

Product Rule: $\boxed{p(x,y) = p(x|y)p(y) = p(y|x)p(x)}$

More generally:

$$p(x_1, x_2, \ldots, x_d) = p(x_d \mid x_1, \ldots, x_{d-1}) \ldots p(x_3 \mid x_1, x_2)\, p(x_2 \mid x_1)\, p(x_1)$$

Bayes' Rule: $\boxed{p(x|y) = \frac{p(y|x)p(x)}{p(y)}}$

Note: Sometimes the subscripts are not used for marginal and conditional distributions when it is clear from the context

$$p(x,y) = p_{X,Y}(x,y)$$

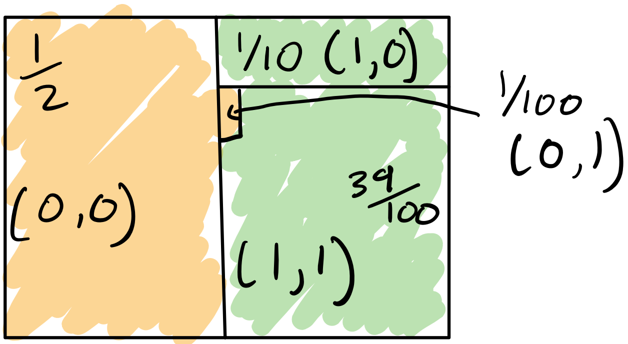
$$p(x) = p_X(x)$$

$$p(y|x) = p_{Y|X}(y|x)$$

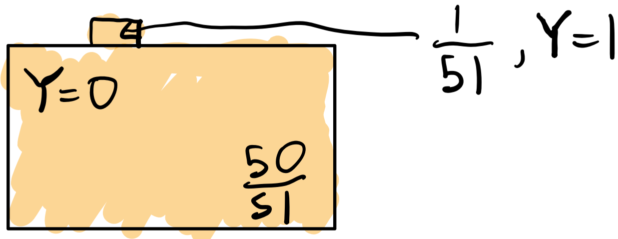
Ex: Probability that I have arthritis given I am young

$$\mathbb{P}(Y=1 \mid X=0) = \sum_{y \in \{1\}} p_{Y|X}(y|0)$$

Arthritis young



condition on being young



$$= P_{Y|X}(1|0)$$

$$= \frac{P_{X,Y}(0,1)}{P_X(0)}$$

$$= \frac{P(0,1)}{P(0,1)+P(0,0)}$$

$$= \frac{1/100}{1/100 + 1/2}$$

$$= \frac{1/100}{51/100} = \frac{1}{51}$$

<u>Ex</u>: Probability of being young given I have arthritis

$$\mathbb{P}(X=0 \mid Y=1) = P_{X|Y}(0|1)$$

Bayes' Rule

$$\overset{\text{Bayes' Rule}}{=} \frac{P_{Y|X}(1|0)\, P_X(0)}{P_Y(1)}$$

$$= \frac{\frac{1}{51} \cdot \frac{51}{100}}{40/100} = \frac{1}{40}$$

<u>Independence</u>: Changing the value of one r.v. doesn't affect the probability of another r.v.

$\boxed{\text{r.v. } X, Y \text{ are independent if: } \quad p(x,y) = p_X(x)\, p_Y(y)}$

Since $p(x,y) = p(x|y)\,p(y) = p(x)\,p(y|x) = p(x)\,p(y)$

$\boxed{\text{independence implies: } p(x|y) = p(x),\ p(y|x) = p(y)}$

More generally:

$X_1, X_2, \ldots, X_d$ are independent if: $p(x_1, \ldots, x_d) = p(x_1) \cdots p(x_d)$

Similarly for distributions:

$\boxed{\text{r.v. } X, Y \text{ are independent if: } \mathbb{P}(X \in E_X, Y \in E_Y) = \mathbb{P}(X \in E_X)\,\mathbb{P}(Y \in E_Y)}$

Ex: X, Y are not independent for Arthritis ex

$$p(0,1) = \frac{1}{100} \neq p_X(0)\, p_Y(1) = \frac{51}{100} \frac{40}{100} = 0.204$$

Ex: $X_1, X_2 \in \{0,1\}$ are flips of two different fair coins

$p(x_1, x_2) = \frac{1}{4}$ for all $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$

$p_{X_1}(x_1)\, p_{X_2}(x_2) = \frac{1}{2} \frac{1}{2} = \frac{1}{4}$

X_1 \ X_2	H	T
H	1/4	1/4
T	1/4	1/4